DOCUMENT RESUME

ED 076 645 TM 002 650

AUTHOR Keesling, J. Ward
TITLE A Problem in the Aggregation of Student Data to the Level of School. Working Draft.
PUB DATE 73
NOTE 7p.; Paper presented at American Educational Research Association Meeting (New Orleans, Louisiana, February 25-March 1, 1973)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Analysis of Covariance; *Data Analysis; Males; *Mathematical Models; *Sampling; Speeches

ABSTRACT

In many circumstances it is appropriate to use the school as the unit of analysis. The variables measured on students must be aggregated to form a mean for each school. However, the means derived from the students sampled in a school will tend to fluctuate around the true mean for the school in a way determined by the within-school correlations among student variables rather than by the between-school correlations. A model is presented which circumvents this problem by obtaining replicate measures for each variable. The model permits estimation of the true between-schools covariance matrix and measurement error variances. An example employing real data is presented. (Author)

FORM 510

PRINTED IN U.S.A.

FILMED FROM BEST AVAILABLE COPY

ED 076645

TM 002 650

WORKING DRAFT

AERA, 1973

A PROBLEM IN THE AGGREGATION OF STUDENT DATA TO THE LEVEL OF SCHOOL

J. Wayne Keesling

Center for the Study of Evaluation
University of California
Los Angeles, California

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

In many circumstances it is appropriate to use the school as the unit of analysis as, for example, in a model comparing treatments affecting entire schools or in a model relating expenditures for instruction to pupil achievement. When the school is the appropriate unit of analysis, the variables measured on students must be aggregated to form a mean (or other measure of central tendency) for each school. When students are sampled in order to generate such a mean, there may be some difficulties in using the aggregated values.

For a particular pair of variables X and Y, the sample of students is supposed to generate a pair of sample means which will be close to the true values for the school. Clearly, there will be some fluctuations from sample to sample; fluctuations which will depend upon the correlation between these variables within the school being sampled. As the between schools correlation of these variables is desired in the analysis, the data are contaminated by the within school correlation when a sampling procedure is used.

In the following example data from each of two samples of boys obtained at each of 39 schools are analysed using a method developed by Joreskog. The analysis shows one way in which the problem of aggregation may be handled.

For each of the samples of boys at each school six variables were measured and the mean values of the observations were computed. Thus each school was represented in the analysis by six pairs of observations -- one mean for each variable from each sample.

Because the values of one replicate measure of all six variables for a school are determined on the same subset of students, the correlations between variables within the replicate measures were systematically higher than the correlations between variables across replicate measures. This implies the presence of a "sampling factor" analogous to the "method factor" in classical test construction (Campbell and Fiske, 1959).

In order to extract the covariance matrix of the latent or "true" variables of interest, the method of analysis of covariance structures (Joreskog, 1970) was applied to these data. Two models may be entertained to account for the observed covariance matrix of 12 variables. The first model posits no "sampling factor" while the second model takes account of this feature of the data explicitly.

Both models may be expressed as parameterizations of the matrix decomposition of the covariance matrix, Σ, given below:

$$\Sigma = \Lambda \Phi \Lambda' + \Psi^2$$

For the model with no sampling factor the parameterization is:

$$\Lambda = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ I & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{matrix} \qquad \bar{\Phi} = \bar{\Phi}$$

$$\Psi = \text{Diagonal}(\psi_1 \psi_2 \psi_3 \psi_4 \psi_5 \psi_6 \psi_1 \psi_2 \psi_3 \psi_4 \psi_5 \psi_6)$$

The ones in the matrix Λ indicate that the solution is restricted to the value 1.0 in that location. The reason for this restriction is provided in the discussion of the alternative model whose parameterization is:

$$\bar{\Phi} = \begin{bmatrix} \Theta & 0' \\ 0 & I \end{bmatrix}$$

$$\Psi = \text{Diagonal}(\psi_1 \psi_2 \psi_3 \psi_4 \psi_5 \psi_6 \psi_1 \psi_2 \psi_3 \psi_4 \psi_5 \psi_6) .$$

$$\Lambda = \begin{matrix} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \lambda_1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \lambda_3 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \lambda_4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \lambda_5 & 0 \\ I & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \lambda_1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \lambda_2 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \lambda_3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \lambda_4 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \lambda_5 \end{matrix}$$

Where a one appears in the above matrices, the solution is restricted to the value 1.0 for that location. Zeroes represent locations restricted to the value 0.0 in the solution. The matrix Λ has six columns for the trait factors in which the loadings are restricted to 1.0 . This follows from the sampling structure for the replicate measures. Each measure consists of the same items, merely different persons from the same school. It seems appropriate to restrict the loading of each measure on the underlying true variable to be 1.0 . The next two columns of Λ represent sampling factors. In order to make the parameters associated with the sampling factors identifiable, the leading parameter must be restricted to have a weight of 1.0 on each sampling factor and the variance-covariance matrix of sampling factors must be the identity matrix. In addition, the elements of the sampling factors which are to be estimated are restricted to be equal for both replicate measures due to the sampling nature of the replicate measures.

The matrix Φ represents the variance-covariance matrix of the latent (or true) measures. It consists of the submatrices Θ, a 6 x 6 covariance matrix of the true variables of interest; 0, the 2 x 6 matrix of sample factor by trait factor covariances which are restricted to be null; and I, the 2 x 2 identity matrix mentioned above. The estimate of Θ will be used in the causal flow analysis to follow.

The matrix Ψ is also restricted to represent the sampling nature of the replicate measures. The measurement error variances for each replicate measure of a variable are restricted to be equal. Using Joreskog's program for the analysis of covariance structures (1970), the following parameter estimates and standard errors were found for the model with sampling factors:

$\lambda_1 = -0.28 \pm 0.17$ $\quad$ $\psi_1 = 1.03 \pm 0.19$
$\lambda_2 = -0.04 \pm 0.06$ $\quad$ $\psi_2 = 0.91 \pm 0.11$
$\lambda_3 = 0.16 \pm 0.11$ $\quad$ $\psi_3 = 0.32 \pm 0.04$
$\lambda_4 = -1.25 \pm 1.61$ $\quad$ $\psi_4 = 0.62 \pm 0.07$
$\lambda_5 = 0.98 \pm 0.25$ $\quad$ $\psi_5 = 8.81 \pm 1.00$
$\quad$ $\quad$ $\psi_6 = 0.65 \pm 0.26$

$$\Theta = \begin{bmatrix} 3.18 \pm 0.97 & & & & \\ 0.44 \pm 0.46 & 1.44 \pm 0.44 & & & \\ -.98 \pm 0.35 & 0.06 \pm 0.21 & 0.82 \pm 0.20 & & \\ -.44 \pm 0.20 & -.04 \pm 0.13 & 0.20 \pm 0.09 & 0.11 \pm 0.08 & \\ -.82 \pm 4.72 & 3.33 \pm 3.17 & 4.34 \pm 2.22 & 2.25 \pm 1.32 & \\ 1.26 \pm 0.54 & -.03 \pm 0.32 & -.43 \pm 0.22 & -.12 \pm 0.14 & \\ 160.40 \pm 46.20 & & & & \\ 3.27 \pm 3.23 & 1.19 \pm 0.45 & & & \end{bmatrix}$$

The obtained χ^2 of 53.8 on 46 degrees of freedom corresponds to a probability of 0.20 which indicates a reasonably good fit of the model to the data. The model with no sampling factor would not converge properly to a solution. The conclusion which this author draws from this phenomenon is that the sampling factor is quite important to the fit of the model. This further implies that there is a substantial problem in the aggregation of student data to the level of school. In the example the only coefficient of loading on the sampling factor which appears to be significant is for the sixth variable. Presumedly; this means that the first and sixth variables are most responsible for the existance of the aggragation difficulty. Substantively, this makes sense as well for the first variable is "father's education (in years)" and the sixth variable is "obtained test score". The correlation of these two variables is well documented.

One caution may be put forward in the recommendation of this method. A similar analysis was performed for girls, and while the general result (lack of convergence for the no sample factor model; reasonably good fit for the alternative model) was the same, the estimated matrix Θ (the estimated true covariance matrix of the variables of interest which would be used in subsequent analyses) proved to have a negative latent root. No good reason can be put forward at this time to explain this ill-conditioned solution.

REFERENCES

Campbell, D. T. and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Joreskog, K. G. (1970). A general Method for analysis of covariance structures. Biometrika, 57, 239-251.